

Next Token Prediction

Nipun Batra

IIT Gandhinagar

August 1, 2025

NEXT TOKEN GENERATION

— Inspired by great lecture(s)

from Andrej Karpathy.

↳ Search for

Neural Networks

Zero to Hero

— We discussed relevance to chat GPT

a b b _ ?

What is the next character?

a b b - ?

What is the next character?

Pose as classification task

a b b

x	$P(x)$
a	0.01
b	...
⋮	
z	0.4
⋮	
-	

Specific Problem

- Generate Indian names
- Dataset :
 - aabid
 - aasida
 - aadesh
 - .
 - :
 - '
 - .
 - zeel

Specific Problem

- Generate Indian names

- Dataset :

- aabid
- aasida
- aadesh
- ⋮
- ⋮
- ⋮
- ⋮
- zeel

Assume

1) Only 26 lower case char

2) - indicates end char

3) $4 < \text{len} < 10$

Generate Training Dataset

WORD #1

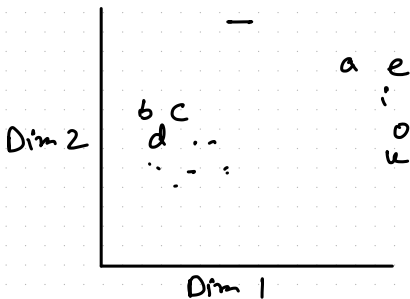
aabid

say we consider			'history / content' of 3 chars.
X			y
-	-	-	a
-	-	a	a
-	a	a	b
a	a	b	i
a	b	i	d
b	i	d	-

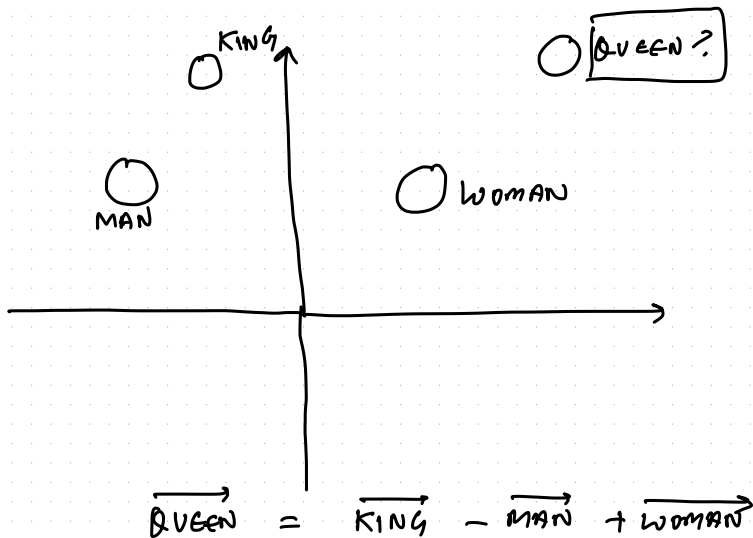
7 training
examples from
1 name

Important Idea Representation

- learn a vector representation for each character
- 'Similar' characters → closer in vector space



WORD2VEC





CHILD
CRYING

=



CHILD
SMILING

+



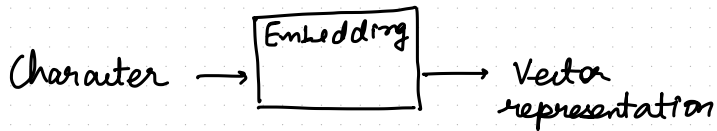
ADULT
CRYING

-

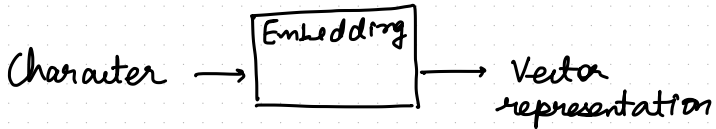


ADULT
SMILING

Embedding matrix | table



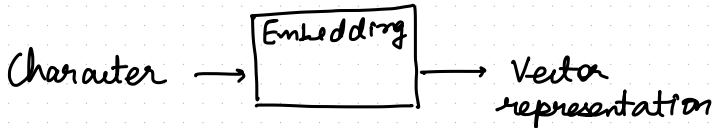
Embedding matrix | table



— Given 27 char (a, .. z, -), 'k' dim embedding

	Dim 1	Dim 2 ..	Dim k
a	0.1
b			
⋮			
i			
z			
-			

Embedding matrix | table



— Given 27 char (a, .. z, -), 'k' dim embedding

	Dim1	Dim2 ..	Dim k
a	0.1
b			
⋮			
i			
z			
-			

\leftarrow **LEARNABLE**

OVERALL ARCHITECTURE

- For illustratⁿ, 2dim embedding
x

a	b	i
---	---	---

OVERALL ARCHITECTURE

i) LOOKUP EMBEDDING

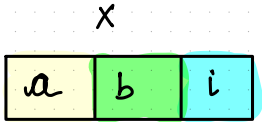
x

a	b	i
---	---	---

	D_1	D_2
a	0.1	0.3
b	-0.1	0.1
:		
:		
i	0.6	0.4
z		
-		

OVERALL ARCHITECTURE

i) LOOKUP EMBEDDING



	D_1	D_2
a	0.1	0.3
b	-0.1	0.1
⋮		
i	0.6	0.4
z		
-		

OVERALL ARCHITECTURE

2) CONCATENATE EMBEDDINGS

X

a	b	i
---	---	---

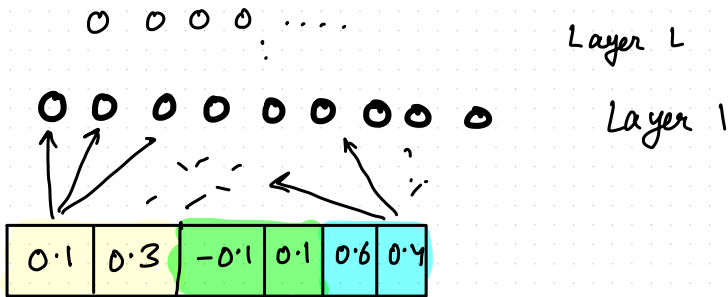
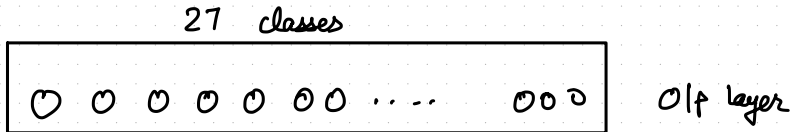
Feature Vector

0.1	0.3	-0.1	0.1	0.6	0.4
-----	-----	------	-----	-----	-----

	D_1	D_2
a	0.1	0.3
b	-0.1	0.1
:		
:		
i	0.6	0.4
z		
-		

OVERALL ARCHITECTURE

3) MLP

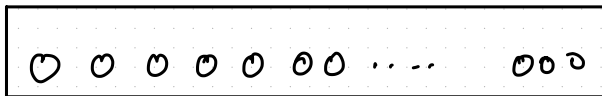


OVERALL ARCHITECTURE

4) USE CROSS ENTROPY LOSS TO LEARN

- 1) Embeddings
- 2) MLP weights

27 classes

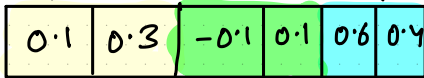


0 0 0 0 : ...

Layer L

0 0 0 0 0 0 0 0 0

Layer 1

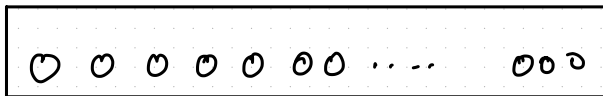


OVERALL ARCHITECTURE

4) USE CROSS ENTROPY LOSS TO LEARN

27 classes

- 1) Embeddings
- 2) MLP weights

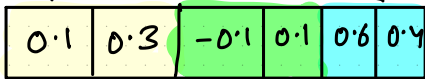


0 0 0 0 : ...

Layer L

0 0 0 0 0 0 0 0 0

Layer 1



OVERALL ARCHITECTURE

5) GENERATION / SAMPLING

Test $i|f$

a b i

Test o/p

c	$P(c)$
a	0.01
b	0.01
c	0.01
d	0.6
e	:
.	,
.	.
.	.
.	.
.	.
.	.
z	\
-	:

OVERALL ARCHITECTURE

5) GENERATION / SAMPLING

Test $i|P$

a b i

Sample from Prob. distribution

ab i a 1.1.

$$ab^i b \quad (i).$$

abid 607.

•

Test o/p

c	$P(c)$
a	0.01
b	0.01
c	0.01
d	0.6
e	:
.	,
.	.
.	:
.	:
.	\
z	:
-	:

OVERALL ARCHITECTURE

5) GENERATION / SAMPLING

